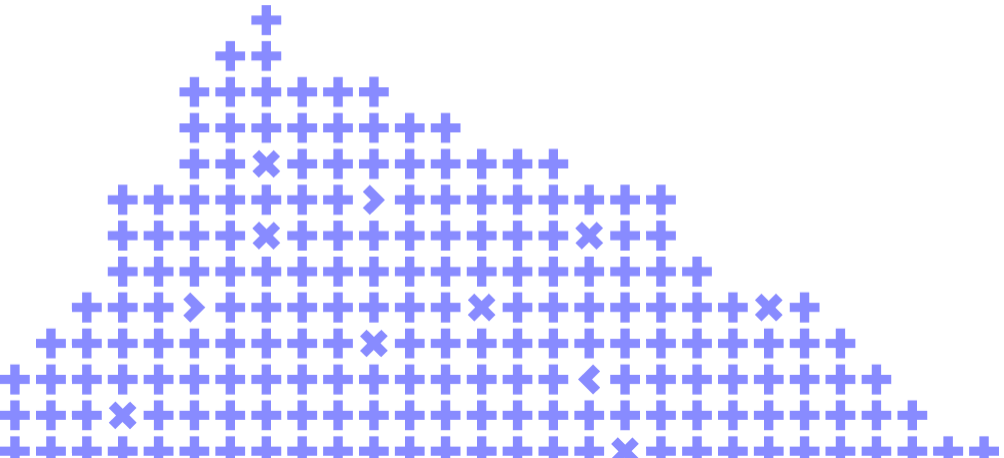
How to create a fully transparent MongoDB database cluster holding terabytes of data serving hundreds of millions of users simultaneously

Arshak Matevosyan



High Load Armenia

Co-organizer

Yandex

Who are we?

A creative platform where people can

create

remix

share

photos

videos

One of the largest open-source content collections in the world

The app is available in 30 languages

World's largest digital creative platform and a top 20 most downloaded app

Picsart milestones (so far)

- 2011 year founded
- 800+ edits every second
- 1B+ edits per month
- 180 countries with active Picsart creators
- 150M+ monthly active users
- 30 supported languages

Make your milestone

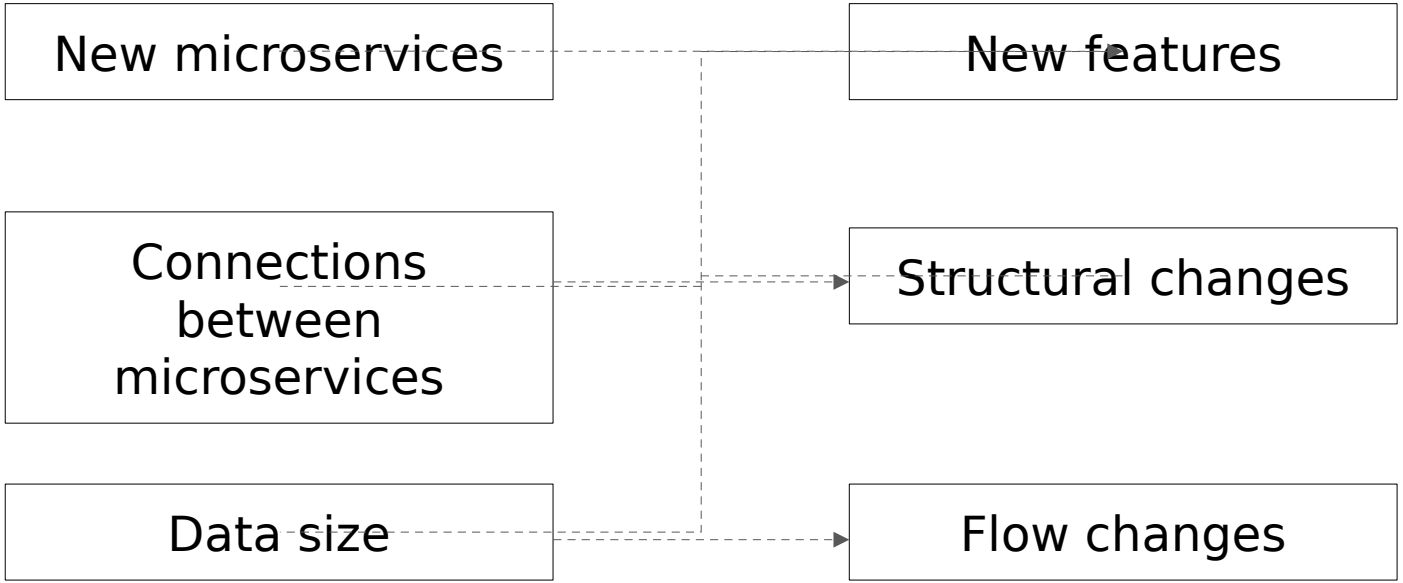
Long story short

The reason behind the creation of Picsart?

Did Picsart change, or no?

What to expect in the future?

Implications of the changes



Challenges

Do we have time to analyze our job result?

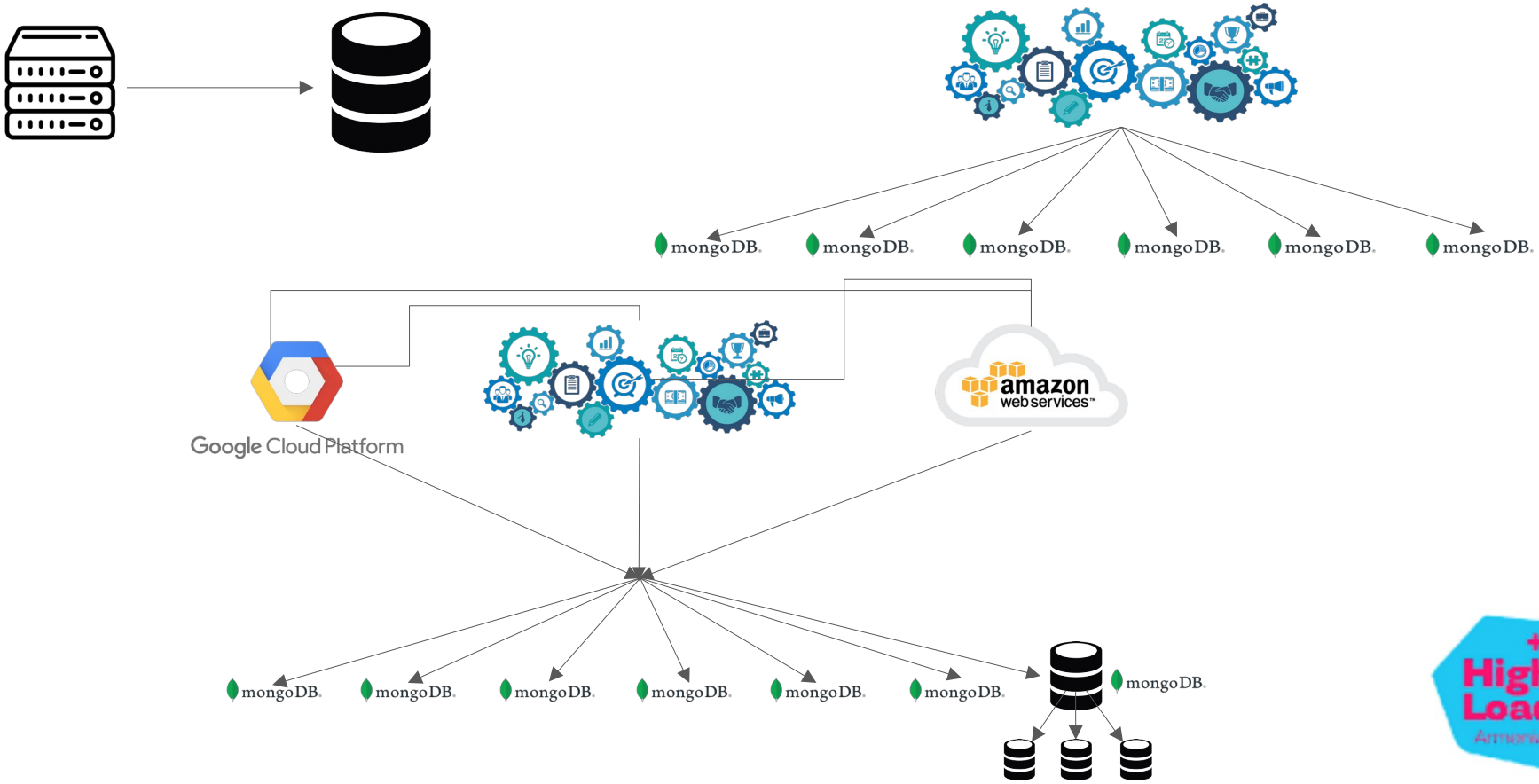
Should we still work reactively? Or go proactive?

How can we predict issues?

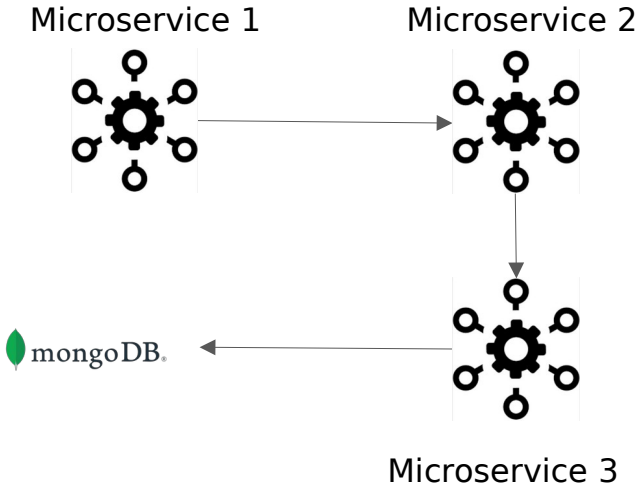
How we can prevent issues?

What will we learn after every incident / issue?

Data flow

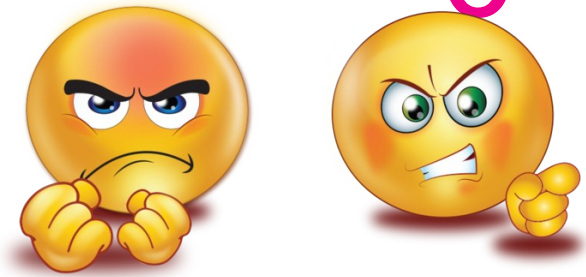

mongoDB.
mongoDB.
mongoDB.
mongoDB.
mongoDB.
mongoDB.
Google Cloud Platform
amazon web services
mongoDB.
mongoDB.
mongoDB.
mongoDB.
mongoDB.
mongoDB.
mongoDB.

Query lifecycle issues



Query lifecycle

Should take 1 0 x ms



What's the core problem

Easy way - Blame each other

Hard way - Work together

Engineers are engineers

Plan together

Work together

Fix all issues together

Party together too

Or be ready for a disaster

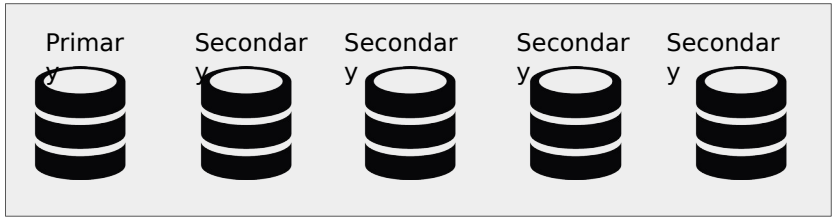
Database queries

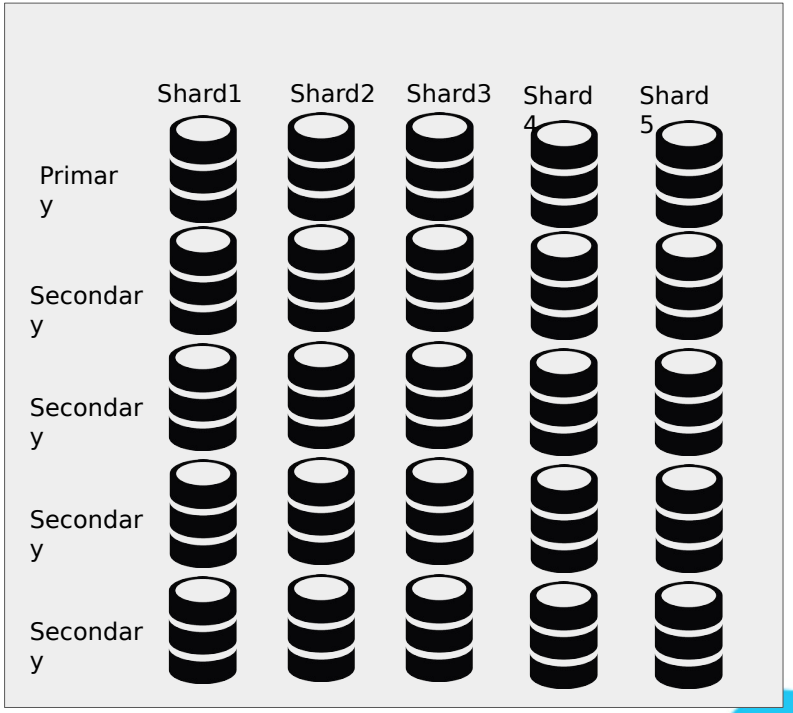
1. Collect all queries from databases

2. Temporary store them somewhere

3. Filter all sensitive and unnecessary data
4. Move to search and analytics engine

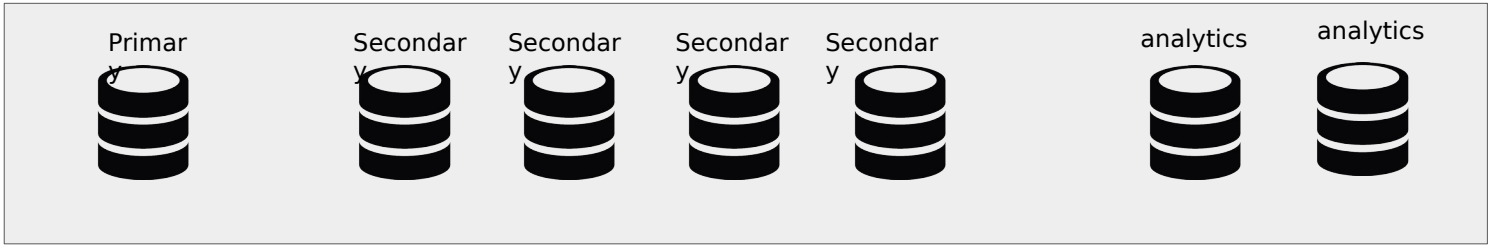
5. Create dashboards





1. Collect all queries from databases

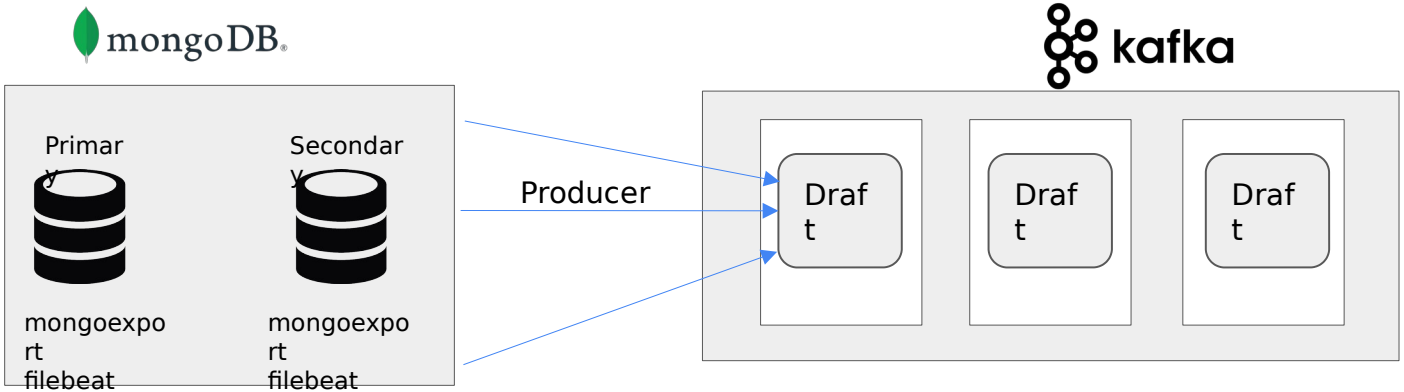
Replication factor	4-9 nodes
Primary, read replicas, analytics replicas	1 master, 2-6 read replicas, 1-2 analytics replicas
How to keep query consistency?	Collect all queries from all nodes, except analytics



1. Collect all queries from databases

Cont'd

How to collect the data right way?	Kafka connector or mongoexport?
How to read all queries and where from?	mongoexport from all nodes
Where to transfer them?	Kafka OnPrem is the best option
How to transfer them?	Filebeat



2. Temporary store them
3. Filter all sensitive data

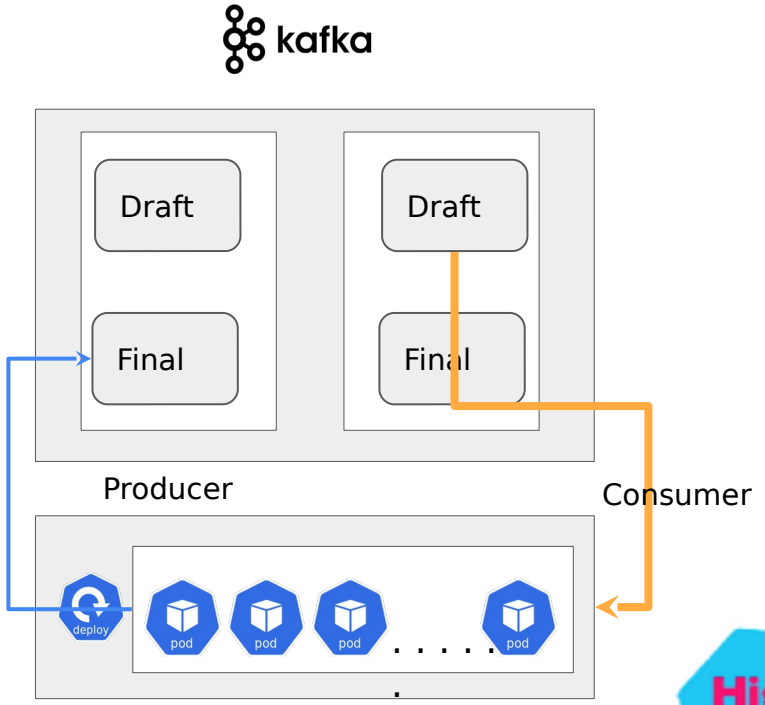
Keep all data in kafka draft topic

Prepare a deployment to filter data

~~Autoscale deployment~~

Filter data and put in final topic

Analyze data and alert



4. Move to search and analytics engine
5. Create dashboards

Transfer all data in ES via logstash

Create dashboards to have visibility

Give access to all relevant engineers

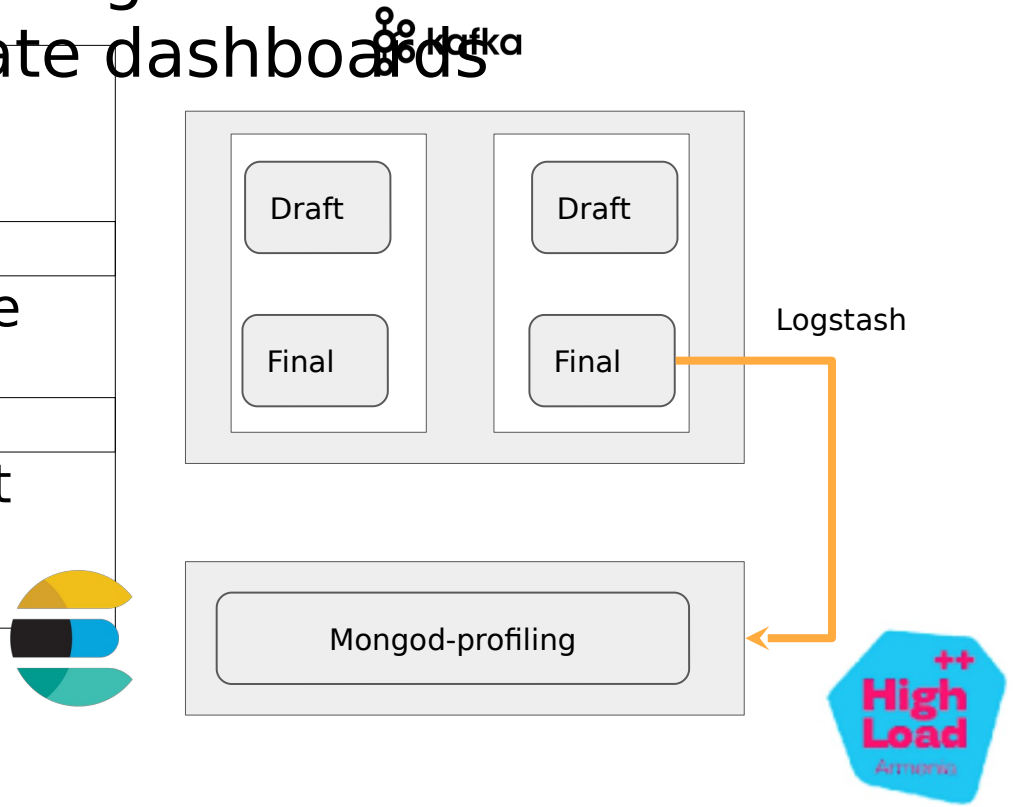
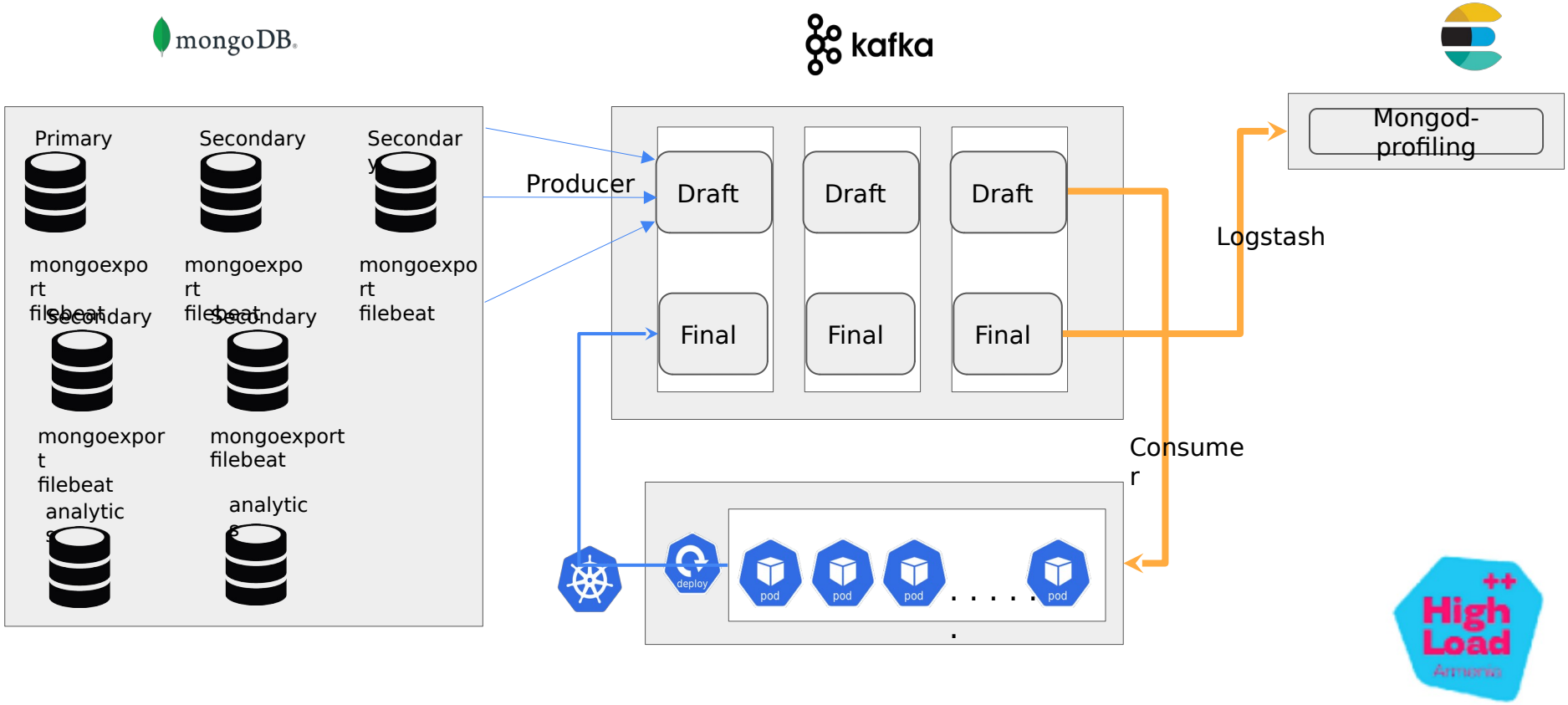
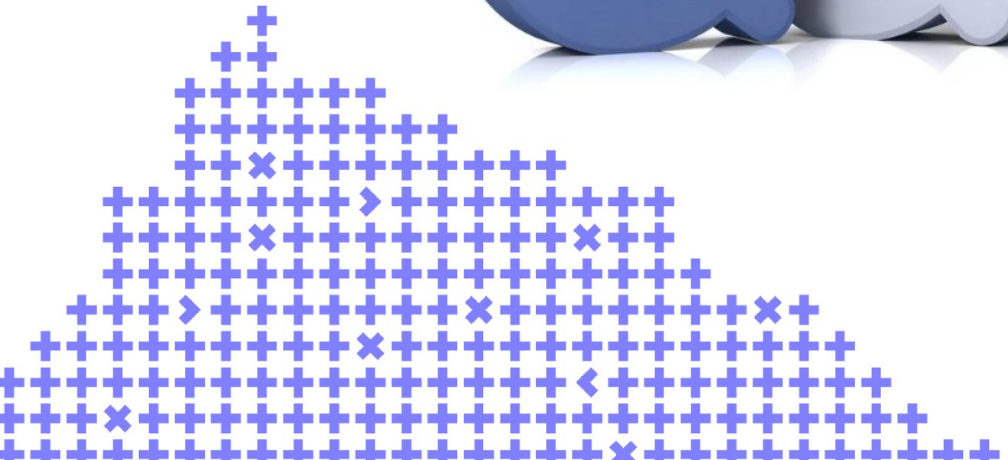


Diagram of query collection



Arshak Matevosyan

Picsart Inc.



Co-organizer

Yandex